Cambridge Reference Sequence

A group under Dr. Fred Sanger at Cambridge University sequenced the mitochondrial genome of one individual of European descent during the 1970s, determining it to have a length of 16,569 base pairs (0.0006% of the total human genome) containing some 37 genes. The *Cambridge Reference Sequence* (CRS) for human mitochondrial DNA was first published in 1981 leading to the initiation of the human genome project.

When other researchers repeated the sequencing, some striking discrepancies were noted. The original published sequence included eleven errors, including one extra base pair in position 3107, and incorrect assignments of single base pairs. Some of these were the result of contamination with bovine and HeLa specimens. The corrected revised CRS was published by Andrews et al. in 1999 and is designated as rCRS.

When mitochondrial DNA sequencing is used for genealogical purposes, the results are often reported as differences from the revised CRS. This notation form is illustrated in the example below, in which a fictional reference sequence is used.

reference: GCTGTCCAGATA

sequence: GCTCTCTAGAGA $\longrightarrow$ 4C,7T,11G

In this notation, `4C` indicates that the sequence differs from the reference sequence at the fourth position, in the sense that there the base `C` is present (whereas at the corresponding position in the reference sequence, the base is `G`). In exactly the same way `7T` indicates that the sequence at the seventh position differs from the reference sequence, because there the base is `T` (whereas at the corresponding position in the reference sequence, the base is `C`). Observed differences to the reference sequence are separated by a comma. If the sequence does not differ from the reference sequence at any position, this is written as an empty string.

Assignment

1. Write a function `seq2dif`, which returns the observed differences between a given sequence `seq` and a given reference sequence `refseq` in the format which has been explained above. Both sequences must be passed to the function as a parameter.
2. Write a function `dif2seq` that returns the original sequence when the observed differences `diff` and the reference sequence `refseq` is given. A string with the observed differences and the reference sequence should be passed to the function as a parameter.

Example

```
>>> seq2diff(seq='GCTCTCTAGAGA', refseq='GCTGTCCAGATA')
'4C,7T,11G'
>>> diff2seq(diff='4C,7T,11G', refseq='GCTGTCCAGATA')
'GCTCTCTAGAGA'
```

Onder leiding van Dr. Fred Sanger sequeneerde een groep onderzoekers van de Universiteit van Cambridge in de jaren '70 het mitochondriaal genoom (mtDNA) van een individu met Europese afkomst. De totale lengte van deze sequentie bedroeg 16.568 baseparen (0.0006%

van het volledig menselijk genoom) waarin 37 genen waren gecodeerd. Deze sequentie werd voor het eerst gepubliceerd in 1981, en wordt algemeen aangeduid als de *Cambridge Reference Sequence* (CRS). De CRS wordt algemeen beschouwd als de eerste stap tot het in kaart brengen van het volledig menselijk genoom.

Toen andere onderzoekers ook mitochondriaal begonnen te sequeneren, doken enkele opvallende discrepanties op. De origineel gepubliceerde sequentie bleek niet minder dan 11 fouten te bevatten, waaronder één extra basepaar op positie 3107 en enkele foutieve bepalingen van individuele baseparen. Enkele van deze fouten waren het resultaat van contaminaties met celmateriaal van runderachtigen en HeLa cellen (onsterfelijke cellijnen die vaak worden gebruikt in wetenschappelijk onderzoek). Daarom publiceerde Richard Andrews in 1999 een gecorrigeerde versie van de referentiesequentie van het menselijk mtDNA, die wordt aangeduid als de revised CRS of kortweg rCRS.

Wanneer mitchondriaal DNA wordt geanalyseerd voor genealogische studies, worden de resultaten vaak gerapporteerd als waargenomen verschillen tegenover de rCRS. Deze notatievorm wordt geïllustreerd in onderstaand voorbeeld, waarbij gebruik wordt gemaakt van een fictieve referentiesequentie.

referentie: GCTGTCCAGATA

sequentie: GCTCTCTAGAGA $\longrightarrow$ 4C,7T,11G

In deze notatie geeft `4C` aan dat de sequentie op de vierde positie verschilt van de referentiesequentie, in die zin dat daar de base `C` staat (waar op de overeenkomstige positie in de referentiesequentie de base `G` staat). Op precies dezelfde manier geeft `7T` aan dat de sequentie op de zevende positie verschilt van de referentiesequentie omdat daar de base `T` staat (waar op de overeenkomstige positie in de referentiesequentie de base `C` staat). Waargenomen verschillen ten overstaan van de referentiesequentie worden van elkaar gescheiden door een komma. Indien de sequentie op geen enkele positie verschilt van de referentiesequentie, dan wordt dit genoteerd als een lege string.

Opgave

1. Schrijf een functie `seq2dif`, die voor een gegeven sequentie `seq` en een gegeven referentiesequentie `refseq` de waargenomen verschillen tegenover de referentiesequentie teruggeeft in het formaat dat hierboven werd toegelicht. Beide sequenties moeten als parameter aan de functie doorgegeven worden.
2. Schrijf een functie `dif2seq`, die voor een reeks waargenomen verschillen `diff` en een gegeven referentiesequentie `refseq` de originele sequentie als resultaat teruggeeft. Een string met de waargenomen verschillen en de referentiesequentie moeten als parameter aan de functie doorgegeven worden.

Voorbeeld

```
>>> seq2diff(seq='GCTCTCTAGAGA', refseq='GCTGTCCAGATA')
'4C,7T,11G'
>>> diff2seq(diff='4C,7T,11G', refseq='GCTGTCCAGATA')
'GCTCTCTAGAGA'
```